C. A. Trugenberger
InfoCodex SA, av. Louis-Casai 18, CH-1209 Geneva, Switzerland
Theory Division, CERN, CH-1211 Geneva 23, Switzerland
e-mail: ca.trugenberger@InfoCodex.com
(October 28, 2002)

I review and expand the model of *quantum associative memory* that I have recently proposed. In this model binary patterns of n bits are stored in the quantum superposition of the appropriate subset of the computational basis of n qbits. Information can be retrieved by performing an input-dependent rotation of the memory quantum state within this subset. The amplitudes of the rotated memory state are peaked on those stored patterns which are closest in Hamming distance to the input. The accuracy of pattern recall can be tuned by adjusting a parameter playing the role of an effective temperature. This model solves the well-known capacity shortage problem of classical associative memories, providing an *exponential improvement* in capacity. The price to pay is the probabilistic nature of information retrieval, a feature that, however, this model shares with our own brain.

I. INTRODUCTION

The power of quantum computation [1] is mostly associated with the speed-up in computing time it can provide with respect to its classical counterparts, the paramount examples being Shor's factoring algorithm [2] and Grover's search algorithm [3]. There is, however, another aspect of quantum computation which represents a big improvement upon its classical counterpart [4]. This leads to an exponential increase in a particular memory capacity rather than speed. In this paper I will review and expand the main aspects of this new application of quantum information theory. Further aspects of it can be found in [5].

In traditional computers the storage of information requires setting up a lookup table (RAM). The main disadvantage of this address-oriented memory system lies in its rigidity. Retrieval of information requires a precise knowledge of the memory address and, therefore, incomplete or corrupted inputs are not permitted.

This is definitely not how our own brain works. When trying to recognize a person from a blurred photo it is totally useless to know that it is the 17384th person you met in your life. Rather, the recognition process is based on our strong power of association with stored memories that resemble the given picture. Association is what we use every time we solve a crossword puzzle and is distinctive of the human brain.

Given the superior power of associative pattern recognition for complex tasks, the shortcomings of RAM memories were addressed by introducing models of associative (or content-addressable) memories [6]. Here, recall of information is possible on the basis of partial knowledge of their content, without knowing the storage location. These are examples of collective computation on neural networks [6], the best known example being the Hopfield model [7] and its generalization to a bidirectional associative memory [8].

While these models solve the problem of recalling incomplete or noisy inputs, they suffer from a severe capacity shortage. Due to the phenomenon of crosstalk, which is essentially a manifestation of the spin glass transition [9] in the corresponding spin systems, the maximum number of binary patterns that can be stored in a Hopfield network of n neurons is $p_{max} \simeq 0.14\, n$ [6] . While various possible improvements can be introduced [6], the maximum number of patterns remains linear in the number of neurons, $p_{max} = O(n)$.

Quantum mechanics offers a way out from the impossibility of reconciling the association power of content-addressable memories with the requirement of large storage capacity. Indeed, quantum mechanical entanglement provides a natural mechanism for both improving dramatically the storage capacity of associative memories and retrieving corrupted or incomplete information.

The basic idea is to store the given p binary patterns of n bits in a quantum superposition of the corresponding subset of the computational basis of n qbits. The number of binary patterns that can be stored in such a *quantum associative memory* is exponential in the number n of qbits, $p_{max} = 2^n$, i.e. it is optimal in the sense that all binary patterns that can be formed with n bits can be stored.

The basic idea of the information retrieval mechanism is very simple. Given an input pattern, the memory quantum state is rotated within the subspace defined by the stored patterns so that the resulting amplitudes are peaked on the stored patterns which are closest in Hamming distance to the input. A measurement of the rotated memory quantum state provides the output pattern.

An efficient way to perform this rotation is to embed the memory quantum state in a larger Hilbert space by adding b control qbits. The full state is then rotated in the enlarged Hilbert space. After this rotation one is interested only in the projection of the rotated state onto a specific subspace of the enlarged Hilbert space. This projection can be obtained either by repeated measurement or by rotating the state (approximately) to the

desired subspace using the amplitude amplification technique [10]. Either way one has to repeat a certain algorithm a number of times and measure the control register to check if the desired projection has been obtained. The information retrieval mechanism is thus probabilistic, with postselection of the measurement result. This means that one has to repeat an algorithm until a threshold T is reached or the measurement of a control register yields a given result. In the former case the input is not recognized. In the latter case, instead, the output is determined itself by a probability distribution on the memory which is peaked around the stored patterns closest in Hamming distance to the input.

The accuracy of this information retrieval mechanism depends on the distribution of the stored patterns. Recognition efficiency is best when the number of stored patterns is very large while identification efficiency is best for isolated patterns which are very different from all other ones, both very intuitive features. Both efficiencies can be tuned to prescribed accuracy levels. The recognition efficiency can be varied by changing the threshold T: the higher T, the larger the number of qbits that can be corrupted without affecting recognition. The identification efficiency, instead, can be tuned by varying the number b of control qbits in the memory. As we shall see, $b = 1/t$ plays the role of an inverse effective temperature t. The lower t, the more concentrated is the corresponding effective Boltzmann distribution on the states closest (in Hamming distance) to the input and the better becomes the identification.

By averaging over the distribution of stored patterns one can eliminate the dependence on the stored pattern distribution and derive the effective statistical mechanics of quantum associative memories by introducing the usual thermodynamic potentials. In particular, the free energy $F(t)$ describes the average behaviour of the recall mechanism at temperature t and provides concrete criteria to tune the accuracy of the associative memory. By increasing b (lowering t), the associative memory undergoes a phase transition from a disordered phase with no correlation between input and output to an ordered phase with minimal Hamming distance bewteen the input and the output. This extends to quantum information theory the relation with Ising spin systems known in error-correcting codes [11] and in public key cryptography [12].

II. STORING INFORMATION

Let me start by describing the elementary quantum gates [1] that I will use in the rest of the paper. First of all there are the single-qbit gates NOT, represented by the first Pauli matrix σ_1, and H (Hadamard), with the matrix representation

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} . \tag{1}$$

Then, I will use extensively the two-qbit XOR (exclusive OR) gate, which performs a NOT on the second qbit if and only if the first one is in state $|1\rangle$. In matrix notation this gate is represented as $\mathrm{XOR} = \mathrm{diag}\,(1, \sigma_1)$, where 1 denotes a two-dimensional identity matrix and σ_1 acts on the components $|01\rangle$ and $|11\rangle$ of the Hilbert space. The 2XOR, or Toffoli gate is the three qbit generalization of the XOR gate: it performs a NOT on the third qbit if and only if the first two are both in state $|1\rangle$. In matrix notation it is given by $\mathrm{2XOR} = \mathrm{diag}\,(1, 1, \sigma_1)$. In the storage algorithm I shall make use also of the nXOR generalization of these gates, in which there are n control qbits. This gate is also used in the subroutines implementing the oracles underlying Grover's algorithm [1] and can be realized using unitary maps affecting only few qbits at a time [13], which makes it feasible. All these are standard gates. In addition to them I introduce the two-qbit controlled gates

$$\begin{aligned} CS^i &= |0\rangle\langle 0| \otimes 1 + |1\rangle\langle 1| \otimes S^i \ , \\ S^i &= \begin{pmatrix} \sqrt{\frac{i-1}{i}} & \frac{1}{\sqrt{i}} \\ \frac{-1}{\sqrt{i}} & \sqrt{\frac{i-1}{i}} \end{pmatrix} \ , \end{aligned} \tag{2}$$

for $i = 1, \ldots, p$. These have the matrix notation $CS^i = \mathrm{diag}\left(1, S^i\right)$. For all these gates I shall indicate by subscripts the qbits on which they are applied, the control qbits coming always first.

Given p binary patterns p^i of length n, it is not difficult to imagine how a quantum memory can store them. Indeed, such a memory is naturally provided by the following superposition of n entangled qbits:

$$|m\rangle = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} |p^i\rangle \ . \tag{3}$$

The only real question is how to generate this state unitarily from a simple initial state of n qbits. This can be done as follows.

In constructing $|m\rangle$ I shall use three registers: a first register p of n qbits in which I will subsequently feed the patterns p^i to be stored, a utility register u of two qbits prepared in state $|01\rangle$, and another register m of n qbits to hold the memory. This latter will be initially prepared in state $|0_1, \ldots, 0_n\rangle$. The full initial quantum state is thus

$$|\psi_0^1\rangle = |p_1^1, \ldots p_n^1; 01; 0_1, \ldots, 0_n\rangle \ . \tag{4}$$

The idea of the storage algorithm is to separate this state into two terms, one corresponding to the already stored patterns, and another ready to process a new pattern. These two parts will be distinguished by the state of the second utility qbit u_2: $|0\rangle$ for the stored patterns and $|1\rangle$ for the processing term.

For each pattern p^i to be stored one has to perform the operations described below:

$$|\psi_1^i\rangle = \prod_{j=1}^{n} 2XOR_{p_j^i u_2 m_j} \ |\psi_0^i\rangle \ . \tag{5}$$

This simply copies pattern p^i into the memory register of the processing term, identified by $|u_2\rangle = |1\rangle$.

$$\begin{aligned} |\psi_2^i\rangle &= \prod_{j=1}^{n} NOT_{m_j} \ XOR_{p_j^i m_j} \ |\psi_1^i\rangle \ , \\ |\psi_3^i\rangle &= nXOR_{m_1 \ldots m_n u_1} |\psi_2^i\rangle \ . \end{aligned} \tag{6}$$

The first of these operations makes all qbits of the memory register $|1\rangle$'s when the contents of the pattern and memory registers are identical, which is exactly the case only for the processing term. Together, these two operations change the first utility qbit u_1 of the processing term to a $|1\rangle$, leaving it unchanged for the stored patterns term.

$$|\psi_4^i\rangle = CS_{u_1 u_2}^{p+1-i} \ |\psi_3^i\rangle \ . \tag{7}$$

This is the central operation of the storing algorithm. It separates out the new pattern to be stored, already with the correct normalization factor.

$$\begin{aligned} |\psi_5^i\rangle &= nXOR_{m_1 \ldots m_n u_1} |\psi_4^i\rangle \ , \\ |\psi_6^i\rangle &= \prod_{j=n}^{1} XOR_{p_j^i m_j} \ NOT_{m_j} \ |\psi_5^i\rangle \ . \end{aligned} \tag{8}$$

These two operations are the inverse of eqs.(6) and restore the utility qbit u_1 and the memory register m to their original values. After these operations on has

$$|\psi_6^i\rangle = \frac{1}{\sqrt{p}} \ \sum_{k=1}^{i} |p^i; 00; p^k\rangle + \sqrt{\frac{p-i}{p}} |p^i; 01; p^i\rangle \ . \tag{9}$$

With the last operation,

$$|\psi_7^i\rangle = \prod_{j=n}^{1} 2XOR_{p_j^i u_2 m_j} \ |\psi_6^i\rangle \ , \tag{10}$$

one restores the third register m of the processing term, the second term in eq.(9) above, to its initial value $|0_1, \ldots, 0_n\rangle$. At this point one can load a new pattern into register p and go through the same routine as just described. At the end of the whole process, the m-register is exactly in state $|m\rangle$, eq. (3). Note that, by construction, there are no restrictions on the loading factor p/n.

III. REMEMBERING

A memory is of real value only if it can be used repeatedly. The rules of quantum mechanics, however, imply that, when the memory register is measured to obtain an output, all the information about the entangled superposition of stored patterns is lost. If one does not want to forget everything after the first information retrieval one must therefore keep a master copy of the memory and produce copies out of it when needed.

Unfortunately, this is no simple task since the linearity of quantum mechanics forbids exact universal cloning of quantum states [14]. Universal cloning [15] has two disadvantages: first of all the copies to be used for information retrieval are imperfect, though optimal [16]; secondly, the quality of the master copy decreases with each recall, i.e. the memory is quickly washed out.

This leaves state-dependent cloning as the only viable option. State-dependent cloners are designed to reproduce only a finite number of states and this is definitely enough for our purposes. It is surely possible to use deterministic cloning [14] by embedding the memory state in a set of orthogonal states and construct a cloner for this particular set. Even simpler, however, is to use a probabilistic cloning machine [17]. To this end it is sufficient to consider any dummy state $|d\rangle$ different from $|m\rangle$ (for more than two states the condition would be linear independence) and to construct a probabilistic cloning machine for these two states. This machine reproduces $|m\rangle$ with probability p_m and $|d\rangle$ with probability p_d; a flag tells us exactly when the desired state $|m\rangle$ has been obtained. In order to obtain an exact copy of $|m\rangle$ one needs then $1/p_m$ trials on average. The master copy is exactly preserved.

The cloning efficiencies of the probabilistic cloner of two states are bounded as follows [17]:

$$p_m + p_d \leq \frac{2}{1 + \langle d|m\rangle} \ . \tag{11}$$

This bound can be made large by choosing $|d\rangle$ as nearly orthogonal to $|m\rangle$ as possible. A simple way to achieve this for a large number of patterns is to encode also the state

$$|d\rangle = \frac{1}{\sqrt{p}} \ \sum_{i=1}^{p} (-1)^{i+1} |p^i\rangle \tag{12}$$

together with $|m\rangle$ when storing information. This can be done simply by using alternately the operators S^i and $\left(S^i\right)^{-1}$ in the storing algorithm of section 2. For binary patterns which are all different from one another one has then

$$\begin{aligned} \langle d|m\rangle &= 0 \ , \qquad p \text{ even} \ , \\ \langle d|m\rangle &= \frac{1}{p} \ , \qquad p \text{ odd} \ , \end{aligned} \tag{13}$$

and the bound for the cloning efficiencies is very close to its maximal value 2 in both cases.

The quantum network for the probabilistic cloner of two states has been developed in [18]. It can be constructed exclusively out of the two simple distinguishability tranfer (D) and state separation (S) gates. Note that these gates embody information about the two states to be cloned. Part of the memory, therefore, actually resides in the cloning network.

IV. RETRIEVING INFORMATION

Assume now one is given a binary input i, which might be, e.g. a corrupted version of one of the patterns stored in the memory. The retrieval algorithm requires also three registers. The first register i of n qbits contains the input pattern; the second register m, also of n qbits, contains the memory $|m\rangle$; finally there is a control register c with b qbits all initialized in the state $|0\rangle$.

The full initial quantum state is thus:

$$|\psi_0\rangle = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} |i; p^k; 0_1, \ldots, 0_b\rangle \tag{14}$$

where $|i\rangle = |i_1, \ldots, i_n\rangle$ denotes the input qbits, the second register, m, contains the memory (3) and all b control qbits are in state $|0\rangle$. Applying the Hadamard gate to the first control qbit one obtains

$$\begin{aligned} |\psi_1\rangle = & \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} |i; p^k; 0_1, \ldots, 0_b\rangle \\ & + \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} |i; p^k; 1_1, \ldots, 0_b\rangle \ . \end{aligned} \tag{15}$$

I now apply to this state the following combination of quantum gates:

$$|\psi_2\rangle = \prod_{j=1}^{n} \ NOT_{m_j} \ XOR_{i_j m_j} |\psi_1\rangle \ , \tag{16}$$

As a result of the above operation the memory register qbits are in state $|1\rangle$ if i_j and p_j^k are identical and $|0\rangle$ otherwise:

$$\begin{aligned} |\psi_2\rangle = & \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} |i; d^k; 0_1, \ldots, 0_b\rangle \\ & + \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} |i; d^k; 1_1, \ldots, 0_b\rangle \ , \end{aligned} \tag{17}$$

where $d_j^k = 1$ if and only if $i_j = p_j^k$ and $d_j^k = 0$ otherwise.

Consider now the following Hamiltonian:

$$\begin{aligned} \mathcal{H} &= (d_H)_m \otimes (\sigma_3)_{c_1} \ , \\ (d_H)_m &= \sum_{j=1}^{n} \left(\frac{\sigma_3 + 1}{2} \right)_{m_j} \ , \end{aligned} \tag{18}$$

where σ_3 is the third Pauli matrix. $\mathcal{H}$ measures the number of 0's in register m, with a plus sign if c_1 is in state $|0\rangle$ and a minus sign if c_1 is in state $|1\rangle$. Given how I have prepared the state $|\psi_2\rangle$, this is nothing else than the number of qbits which are different in the input and memory registers i and m. This quantity is called the *Hamming distance* and represents the (squared) Euclidean distance between two binary patterns.

Every term in the superposition (17) is an eigenstate of $\mathcal{H}$ with a different eigenvalue. Applying thus the unitary operator $\exp(i\pi\mathcal{H}/2n)$ to $|\psi_2\rangle$ one obtains

$$|\psi_3\rangle = \mathrm{e}^{i\frac{\pi}{2n}\mathcal{H}} \ |\psi_2\rangle \ , \tag{19}$$

$$\begin{aligned} |\psi_3\rangle = & \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} \mathrm{e}^{i\frac{\pi}{2n} d_H\left(i,p^k\right)} |i; d^k; 0_1, \ldots, 0_b\rangle \\ & + \frac{1}{\sqrt{2p}} \sum_{k=1}^{p} \mathrm{e}^{-i\frac{\pi}{2n} d_H\left(i,p^k\right)} |i; d^k; 1_1, \ldots, 0_b\rangle \ , \end{aligned}$$

where $d_H\left(i, p^k\right)$ denotes the Hamming distance bewteen the input i and the stored pattern p^k.

In the final step I restore the memory gate to the state $|m\rangle$ by applying the inverse transformation to eq. (16) and I apply the Hadamard gate to the control qbit c_1, thereby obtaining

$$|\psi_4\rangle = H_{c_1} \prod_{j=n}^{1} XOR_{i_j m_j} \ NOT_{m_j} \ |\psi_3\rangle \ , \tag{20}$$

$$\begin{aligned} |\psi_4\rangle = & \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \cos \frac{\pi}{2n} d_H\left(i,p^k\right) |i; p^k; 0_1, \ldots, 0_b\rangle \\ & + \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \sin \frac{\pi}{2n} d_H\left(i,p^k\right) |i; p^k; 1_1, \ldots, 0_b\rangle. \end{aligned}$$

The idea is now to repeat the above operations sequentially for all b control qbits c_1 to c_b. This gives

$$\begin{aligned} |\psi_{\mathrm{fin}}\rangle = & \frac{1}{\sqrt{p}} \sum_{k=1}^{p} \sum_{l=0}^{b} \cos^{b-l}\left(\frac{\pi}{2n} d_H\left(i,p^k\right)\right) \times \\ & \sin^{l}\left(\frac{\pi}{2n} d_H\left(i,p^k\right)\right) \sum_{\{J^l\}} |i; p^k; J^l\rangle, \end{aligned} \tag{21}$$

where $\{J^l\}$ denotes the set of all binary numbers of b bits with exactly l bits 1 and $(b-l)$ bits 0. This represents a rotation of the memory quantum state in the enlarged Hilbert space obtained by adding b control qbits.

Note that the overall effect obtained by this rotation is an overall amplitude concentration on memory states similar to the input if there is a large number of $|0\rangle$ control qbits and an amplitude concentration on states different to the input if there is a large number of $|1\rangle$ control qbits. As a consequence, the most interesting state for information retrieval purposes is the projection of $|\psi_{\mathrm{fin}}\rangle$ onto the subspace with all control qbits in state $|0\rangle$. There are two ways of obtaining this projection. The first is to repeat the above algorithm and measure the control register several times, until exactly the desired state for the control register is obtained. If the number of such repetitions exceeds a preset threshold T the input is classified as "non-recognized" and the algorithm is stopped. Otherwise, once $|c_1, \ldots, c_b\rangle = |0_1, \ldots, 0_b\rangle$ is obtained, one proceeds to a measurement of the memory register m, which yields the output pattern of the memory.

The second method is to apply T steps of the amplitude amplification algorithm [10] rotating $|\psi_{\rm fin}\rangle$ towards its projection onto the "good" subspace formed by the states with all control qbits in state $|0\rangle$. If a measurement of the control register yields $|0_1 \dots 0_b\rangle$ one proceeds as before to a measurement of the memory register; otherwise the input is classified as "non-recognized".

Since the expected number of repetitions needed to measure the desired control register state is $1/P_b^{\rm rec}$, with

$$P_b^{\rm rec} = \frac{1}{p} \sum_{k=1}^{p} \cos^{2b}\left(\frac{\pi}{2n} d_H\left(i; p^k\right)\right) \tag{22}$$

the probability of measuring $|c_1, \dots, c_n\rangle = |0_1, \dots, 0_n\rangle$, the threshold T governs the *recognition efficiency* of the input patterns. Note, however, that amplitude amplification provides a quadratic boost [10] to the recognition efficiency since only $1/\sqrt{P_b^{\rm rec}}$ steps are required to rotate $|\psi_{\rm fin}\rangle$ onto the desired subspace. Accordingly, the threshold T can be lowered to $\sqrt{T}$ with respect to the method of projection by measurement.

Once the input pattern i is recognized, the measurement of the memory register yields the stored pattern p^k with probability

$$P_b\left(p^k\right) = \frac{1}{Z} \cos^{2b}\left(\frac{\pi}{2n} d_H\left(i, p^k\right)\right) \ , \tag{23}$$

$$Z = p P_b^{\rm rec} = \sum_{k=1}^{p} \cos^{2b}\left(\frac{\pi}{2n} d_H\left(i, p^k\right)\right) \ . \tag{24}$$

Clearly, this probability is peaked around those patterns which have the smallest Hamming distance to the input. The highest probability of retrieval is thus realized for that pattern which is most similar to the input. There are no restrictions on the loading factor p/n coming from the information retrieval algorithm.

In addition to the threshold T, there is a second tunable parameter, namely the number b of control qbits. This new parameter b controls the *identification efficiency* of the quantum memory since, increasing b, the probability distribution $P_b\left(p^k\right)$ becomes more and more peaked on the low $d_H\left(i, p^k\right)$ states, until

$$\lim_{b\to\infty} P_b\left(p^k\right) = \delta_{k k_{\rm min}} \ , \tag{25}$$

where $k_{\rm min}$ is the index of the pattern (assumed unique for convenience) with the smallest Hamming distance to the input.

The probability of recognition is determined by comparing (even) powers of cosines and sines of the distances to the stored patterns. It is thus clear that the worst case for recognition is the situation in which there is an isolated pattern, with the remaining patterns forming a tight cluster spanning all the largest distances to the first one. As a consequence, the threshold needed to recognize all patterns diminishes when the number of stored patterns becomes very large, since, in this case, the distribution of patterns becomes necessarily more homogeneous. Indeed, for the maximal number of stored patterns $p = 2^n$ one has $P_b^{\rm rec} = 1/2^b$ and the recognition efficiency becomes also maximal, as it should be.

While the recognition efficiency depends on comparing powers of cosines and sines of the same distances in the distribution, the identification efficiency depends on comparing the (even) powers of cosines of the different distances in the distribution. Specifically, it is best when one of the distances is zero, while all others are as large as possible, such that the probability of retrieval is completely peaked on one pattern. As a consequence, the identification efficiency is best when the recognition efficiency is worst and viceversa.

The role of the parameter b becomes familiar upon a closer examination of eq.(23). Indeed, the quantum distribution described by this equation is equivalent to a canonical Boltzmann distribution with (dimensionless) temperature $t = 1/b$ and (dimensionless) energy levels

$$E^k = -2 \log \cos\left(\frac{\pi}{2n} d_H\left(i, p^k\right)\right) \ , \tag{26}$$

with Z playing the role of the partition function.

The appearance of an effective thermal distribution suggests studying the average behaviour of quantum associative memories via the corresponding thermodynamic potentials. Before this can be done, however, one must deal with the different distributions of stored patterns characterizing each individual memory. To this end I propose to average also over this distribution, by keeping as a tunable parameter only the minimal Hamming distance d between the input and the stored patterns. In doing so, one obtains an average description of the average memory. This is essentially the replica trick used to derive the behaviour of spin glasses [9] and classical Hopfield models [6].

As a first step it is useful to normalize the pattern representation by adding (modulo 2) to all patterns, input included, the input pattern i. This clearly preserves all Hamming distances and has the effect of normalizing the input to be the state with all qbits in state $|0\rangle$. The Hamming distance $d_H\left(i, p^k\right)$ becomes thus simply the number of qbits in pattern p^k with value $|1\rangle$. For loading factors $p/n \to 0$ in the limit $n \to \infty$ the partition function for the average memory takes then a particularly simple form:

$$Z_{\rm av} = \frac{p}{N_\lambda} \sum_{\{\lambda\}} \sum_{j=d}^{n} \lambda_j \cos^{2b}\left(\frac{\pi}{2} \frac{j}{n}\right) \ , \tag{27}$$

where λ_j describes an unconstrained probability distribution such that $\sum_{j=d}^{n} \lambda_j = 1$, $\{\lambda\}$ is the set of such distributions and N_λ the corresponding normalization factor. For finite loading factors, instead, the probabilities λ_j become subject to constraints which make things more complicated.

I now introduce the free energy $F(b,d)$ by the usual definition

$$Z_{\rm av} = p\ {\rm e}^{-bF(b,d)} = Z_{\rm av}(b=0)\ {\rm e}^{-bF(b,d)}\ , \qquad (28)$$

where I have chosen a normalization such that $\exp(-bF)$ describes the deviation of the partition function from its value for $b=0$ (high effective temperature). Since Z/p, and consequently also $Z_{\rm av}/p$ posses a finite, non-vanishing large-n limit, this normalization ensures that $F(b,d)$ is intensive, exactly like the energy levels (26), and scales as a constant for large n. This is the only difference with respect to the familiar situation in statistical mechanics.

The free energy describes the equilibrium of the system at effective temperature $t=1/b$ and has the usual expression in terms of the internal energy U and the entropy S:

$$\begin{aligned} F(t,d) &= U(t,d) - tS(t,d)\ , \\ U(t,d) &= \langle E\rangle_t\ , \quad S(t,d) = \frac{-\partial F(t,d)}{\partial t}\ . \end{aligned} \qquad (29)$$

Note that, with the normalization I have chosen in (28), the entropy S is always a negative quantity describing the deviation from its maximal value $S_{\rm max}=0$ at $t=\infty$.

By inverting eq.(26) with F substituting E one can also define an effective (relative) input/output Hamming distance $\mathcal{D}$ at temperature t:

$$\mathcal{D}(t,d) = \frac{2}{\pi}\ \arccos\ {\rm e}^{\frac{-F(t,d)}{2}}\ . \qquad (30)$$

This corresponds exactly to representing the recognition probability of the average memory as

$$(P_b^{\rm rec})_{\rm av} = \cos^{2b}\left(\frac{\pi}{2}\mathcal{D}(b,d)\right)\ , \qquad (31)$$

which can also be taken as the primary definition of the effective Hamming distance.

The function $\mathcal{D}(b,d)$ provides a complete description of the behaviour of quantum associative memories, which can be used to tune their performance. Indeed, suppose that one wants the memory to recognize and identify inputs with up to ϵn corrupted inputs with an efficiency of ν ($0\le\nu\le 1$). Then one must choose a number b of control qbits sufficiently large that $(\mathcal{D}(b,\epsilon n)-\epsilon)\le(1-\nu)$ and a threshold T of repetitions satisfying $T\ge 1/\cos^{2b}\left(\frac{\pi}{2}\mathcal{D}(b,\epsilon n)\right)$, as illustrated in Fig. 1 below.

A first hint about the general behaviour of the effective distance function $\mathcal{D}(b,d)$ can be obtained by examining closer the energy eigenvalues (26). For small Hamming distance to the input these reduce to

$$E^k \simeq \frac{\pi^2}{4}\left(\frac{d_H\left(i,p^k\right)}{n}\right)^2\ , \qquad \frac{d_H\left(i,p^k\right)}{n} \ll 1\ . \qquad (32)$$

Choosing again the normalization in which $|i\rangle = |0\ldots 0\rangle$ and introducing a "spin" s_i^k with value $s_i^k=-1/2$ if qbit i in pattern p^k has value $|0\rangle$ and $s_i^k=+1/2$ if qbit i in pattern p^k has value $|1\rangle$, one can express the energy levels for $d_H/n\ll 1$ as

$$E^k = \frac{\pi^2}{16} + \frac{\pi^2}{4n^2}\sum_{i,j} s_i^k s_j^k + \frac{\pi^2}{4n}\sum_i s_i^k\ . \qquad (33)$$

Apart from a constant, this is the Hamiltonian of an infinite-range antiferromagnetic Ising model in presence of a magnetic field. The antiferromagnetic term favours configurations k with half the spins up and half down, so that $s^k_{\rm tot}=\sum_i s_i^k = 0$, giving $E^k=\pi^2/16$.The magnetic field, however, tends to align the spins so that $s^k_{\rm tot}=-n/2$, giving $E^k=0$. Since this is lower than $\pi^2/16$, the ground state configuration is ferromagnetic, with all qbits having value $|0\rangle$. At very low temperature (high b), where the energy term dominates the free energy, one expects thus an ordered phase of the quantum associative memory with $\mathcal{D}(t,d)=d/n$. This corresponds to a perfect identification of the presented input. As the temperature is raised (b decreased) however, the thermal energy embodied by the entropy term in the free energy begins to counteract the magnetic field. At very high temperatures (low b) the entropy approaches its maximal value $S(t=\infty)=0$ (with the normalization chosen here). If this value is approached faster than $1/t$, the free energy will again be dominated by the internal energy . In this case, however, this is not any more determined by the ground state but rather equally distributed on all possible states, giving

$$\begin{aligned} F(t=\infty) &= U(t=\infty) = \frac{-1}{1-\frac{d}{n}}\int_{\frac{d}{n}}^1 dx\ 2\ \log\ \cos\left(\frac{\pi}{2}x\right) \\ &= \left(1+\frac{d}{n}\right)2\ \log 2 + O\left(\left(\frac{d}{n}\right)^2\right)\ , \end{aligned} \qquad (34)$$

and leading to an effective distance

$$\mathcal{D}(t=\infty,d) = \frac{2}{3} - \frac{2\ \log 2}{\pi\sqrt{3}}\ \frac{d}{n} + O\left(\left(\frac{d}{n}\right)^2\right)\ . \qquad (35)$$

This value corresponds to a disordered phase with no correlation between input and output of the memory.

A numerical study of the thermodynamic potentials in (29) and (30) indeed confirms a phase transition from the ordered to the disordered phase as the effective temperature is raised. In Fig. 1 I show the effective distance $\mathcal{D}$ and the entropy S for 1 Mb ($n=8\times 10^6$) patterns and $d/n=1\%$ as a function of the inverse temperature b (the entropy is rescaled to the interval [0,1] for ease of presentation). At high temperature there is indeed a disordered phase with $S=S_{\rm max}=0$ and $\mathcal{D}=2/3$. At low temperatures, instead, one is in the ordered phase with $S=S_{\rm min}$ and $\mathcal{D}=d/n=0.01$. The effective Hamming distance plays thus the role of the order parameter for the phase transition.

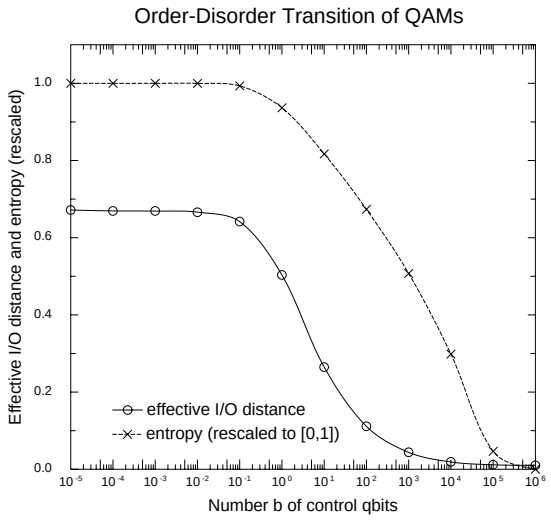

FIG. 1. Effective input/output distance and entropy (rescaled to [0,1]) for 1Mb patterns and $d/n = 1\%$.

The phase transition occurs at $b_{\rm cr} \simeq 10^{-1}$. The physical regime of the quantum associative memory (b = positive integer) begins thus just above this transition. For a good accuracy of pattern recognition one should choose a temperature low enough to be well into the ordered phase.

Having described at length the information retrieval mechanism for complete, but possibly corrupted patterns, it is easy to incorporate also incomplete ones. To this end assume that only $q < n$ qbits of the input are known and let me denote these by the indices $\{k1, \ldots, kq\}$. After assigning the remaining qbits randomly, there are two possibilities. One can just treat the resulting complete input as a noisy one and proceed as above or, better, one can limit the operator $(d_H)_m$ in the Hamiltonian (18) to

$$(d_H)_m = \sum_{i=1}^{q} \left(\frac{\sigma_3 + 1}{2} \right)_{m_{ki}} , \qquad (36)$$

so that the Hamming distances to the stored patterns are computed on the basis of the known qbits only. After this the pattern recall process continues exactly as described above. This second possibility has the advantage that it does not introduce random noise in the similarity measure but it has the disadvantage that the operations of the memory have to be adjusted to the inputs.

V. EFFICIENCY, COMPLEXITY AND MEMORY TUNING

As anticipated in section 4, the effective i/o Hamming distance can be used to tune the quantum associative memory to prescribed accuracy levels. Typically, it is to be expected that increasing this accuracy will lead to an enhanced complexity level. Before I even begin addressing this issue, however, I will show that the information retrieval algorithm is efficient.

First of all I would like to point out that, in addition to the standard NOT, H (Hadamard), XOR, 2XOR (Toffoli) and nXOR gates [1] I have introduced only the two-qbit gates CS^i in eq. (2) and the unitary operator $\exp(i\pi\mathcal{H}/2n)$. This latter can, however also be realized by simple gates involving only one or two qbits. To this end I introduce the single-qbit gate

$$U = \begin{pmatrix} e^{i\frac{\pi}{2n}} & 0 \\ 0 & 1 \end{pmatrix} , \qquad (37)$$

and the two-qbit controlled gate

$$CU^{-2} = |0\rangle\langle 0| \otimes 1 + |1\rangle\langle 1| \otimes U^{-2} . \qquad (38)$$

It is then easy to check that $\exp(i\pi\mathcal{H}/2n)$ in eq. (17) can be realized as follows:

$$e^{i\frac{\pi}{2n}\mathcal{H}} \; |\psi_2\rangle = \prod_{i=1}^{n} \left(CU^{-2}\right)_{cm_i} \; \prod_{j=1}^{n} U_{m_j} \; |\psi_2\rangle , \qquad (39)$$

where c is the control qbit for which one is currently repeating the algorithm. Essentially, this means that one implements first $\exp(i\pi d_H/2n)$ and then one corrects by implementing $\exp(-i\pi d_H/n)$ on that part of the quantum state for which the control qbit $|c\rangle$ is in state $|1\rangle$.

Using this representation for the Hamming distance operator one can count the total number of simple gates that one must apply in order to implement one step of the information retrieval algorithm. This is given by one gate for the operation leading to eq. (15), $2n$ gates for the operation leading to eq. (16), $2n$ gates for the unitary evolution leading to (17) and the again $(2n + 1)$ gates to reverse the first two steps and obtain the final result. This gives a total of $(6n + 2) = (6\,\log(N) + 2)$ gates for each step of the retrieval algorithm, showing that this can be implemented efficiently.

The information retrieval step described above must be repeated for each control qbit, and then one has to repeat the whole process over and over until one obtains the desired state upon measurement of the control register or one reaches a number T of repetitions. The overall complexity C of information retrieval is thus bounded by

$$C \leq Tb(6n + 2) \qquad (40)$$

As expected, this result depends on both T and b, the parameters governing the recognition and identification efficiencies. Indeed it represents exactly the unavoidable tradeoff between accuracy and complexity.

Suppose now one would like to recognize on average inputs with up to 1% of corrupted or missing bits and identify them with high accuracy. The effective i/o Hamming distance $\mathcal{D}$ shown in Fig. 1 can then be used to determine

the values of the required parameters T and b needed to reach this accuracy for the average memory. For $b = 10^4$ e.g., one has $\mathcal{D} = 0.018$, which gives the average i/o distance (in percent of total qbits) if the minimum possible i/o distance is 0.01. For this value of b the recognition probability is $1.5\ 10^{-4}$. With the measurement repetition technique one should thus set the threshold $T = 0.6\ 10^4$. Using amplitude amplification, however, one needs only around $T = 80$ repetitions.

I would like to conclude by stressing that the values of b and T obtained by tuning the memory with the effective i/o Hamming distance become n-independent for large values of n. This is because they are intensive variables unaffected by this "thermodynamic limit". Apart from the linear factor of n determined by a single step of the information retrieval algorithm, the complexity of quantum pattern recognition is thus uniquely determined by the accuracy requirements.

VI. CONCLUSION

I would like to conclude this review by stressing the reason why a quantum associative memory works so much better than its classical counterpart.

In classical associative memories, the information about the patterns to recall is typically stored in an energy functional. When retrieving information, it is the input configuration which evolves to the corresponding output, driven by the memory. The capacity shortage is due to a phase transition in the statistical ensemble governed by the memory energy functional. Spurious metastable minima not associated with any of the original patterns become important for loading factors $p/n > 0.14$ and wash out completely the memory. The important point is that the input/output evolution is governed by a dynamics depending collectively on all stored patterns: this simply becomes unreliable when the degrees of freedom increase too much.

In quantum associative memories, instead, the information is stored on a quantum state. The idea of information retrieval is to perform an input-dependent rotation on this memory state. It is thus the memory itself which evolves towards the correct output, obtained by measurement, while the input is fixed. In this case it is thus the object depending collectively on all stored patterns which evolves, while the dynamics governing this evolution depends essentially only on the input. As a consequence, it is always such that the effective energy functional governing the retrieval process has a unique global minimum on the closest state to the input, without spurious minima that could cause trouble. The price to pay is the probabilistic nature of the information retrieval, caused essentially by all other patterns in the memory. One does not always have the guarantee that an input is recognized, and identified correctly, even when it should, according to the memory tuning. But, after all, the same happens also to the human brain.

[1] For a review see: M. A. Nielsen and I. L. Chuang, "Quantum Computation and Quantum Information", Cambridge University Press, Cambridge (2000); A. O. Pittenger, "An Introdcution to Quantum Computing Algorithms", Birkhäuser, Boston (2000).

[2] P. W. Shor, SIAM J. Computing **26** (1997) 1484.

[3] L. Grover, Phys. Rev. Lett. **79** (1997) 325.

[4] C. A. Trugenberger, Phys. Rev. Lett. **87.067901** (2001), C. A Trugenberger, quant-ph/0200000.

[5] M. Sasaki, A. Carlini and R. Jozsa, Phys. Rev. **A64.022317** (2001); M Sasaki and A. Carlini, quant-ph/0202173; R. Schützhold, quant-ph/0208063.

[6] For a review see: B. Müller and J. Reinhardt, "Neural Networks", Springer-Verlag, Berlin (1990); T. Kohonen, "Self-Organization and Associative Memory", Springer-Verlag, Berlin (1984).

[7] J.J. Hopfield, Proc. Natl. Acad. Scie. USA **79**, 2554 (1982).

[8] B. Kosko, IEEE Trans. on Systems, Man and Cybernetics **18**, 49 (1988).

[9] See e.g. M. Mezard, G. Parisi and M. A. Virasoro, "Spin Glass Theory and Beyond", World Scientific, Singapore (1987).

[10] G. Brassard, P. Hoyer, M. Mosca and A. Tapp, "Amplitude Amplification and Estimation", quant-ph/0005055.

[11] N. Sourlas, Nature **339**, 693 (1989); I. Kanter and D. Saad, Phys. Rev. Lett. **83**, 2660 (1999); Y. Kabashima, T. Murayama and D. Saad, Phys. Rev. Lett. **84**, 1355 (2000).

[12] Y. Kabashima, T. Murayama and D. Saad, Phys. Rev. Lett. **84** 2030 (2000).

[13] See e.g. A. Barenco, C. Bennet, R. Cleve, D. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. Smolin and H. Weinfurter, Phys. Rev. **A52**, 3457 (1995).

[14] W. Wootters and W. Zurek, Nature 299, 802 (1982).

[15] V. Buzek and M. Hillery, Phys. Rev. **A54**, 1844 (1996).

[16] N. Gisin and S. Massar, Phys. Rev. Lett. **79**, 2153 (1997); D. Bruss, A. K. Ekert and C. Macchiavello, Phys. Rev. Lett. **81**, 2598 (1998).

[17] L.-M. Duan and G.-C. Guo, Phys. Rev. Lett. 80, 4999 (1998).

[18] A. Chefles and S. M. Barnett, Phys. Rev. **A60**, 136 (1999).